Data Quality Monitoring using Machine Learning for CMS Experiment at CERN

Guillermo Fidalgo Rodríguez

Physics Department - University of Puerto Rico – Mayagüez

Abstract

The Data Quality Monitoring (DQM) of CMS is a key asset to deliver high-quality data for physics analysis and it is used both in the online and offline environment. The current paradigm of the quality assessment is labor intensive and it is based on the scrutiny of a large number of histograms by detector experts comparing them with a reference. This project aims at applying recent progress in Machine Learning techniques to the automation of the DQM scrutiny. In particular the use of convolutional neural networks to spot problems in the acquired data is presented with particular attention to semi-supervised models (e.g. autoencoders) to define a classification strategy that doesn't assume previous knowledge of failure modes. Real data from the hadron calorimeter of CMS are used to demonstrate the effectiveness of the proposed approach.

What is Data Quality Monitoring (DQM)?

Two kinds of workflows:

Online DQM

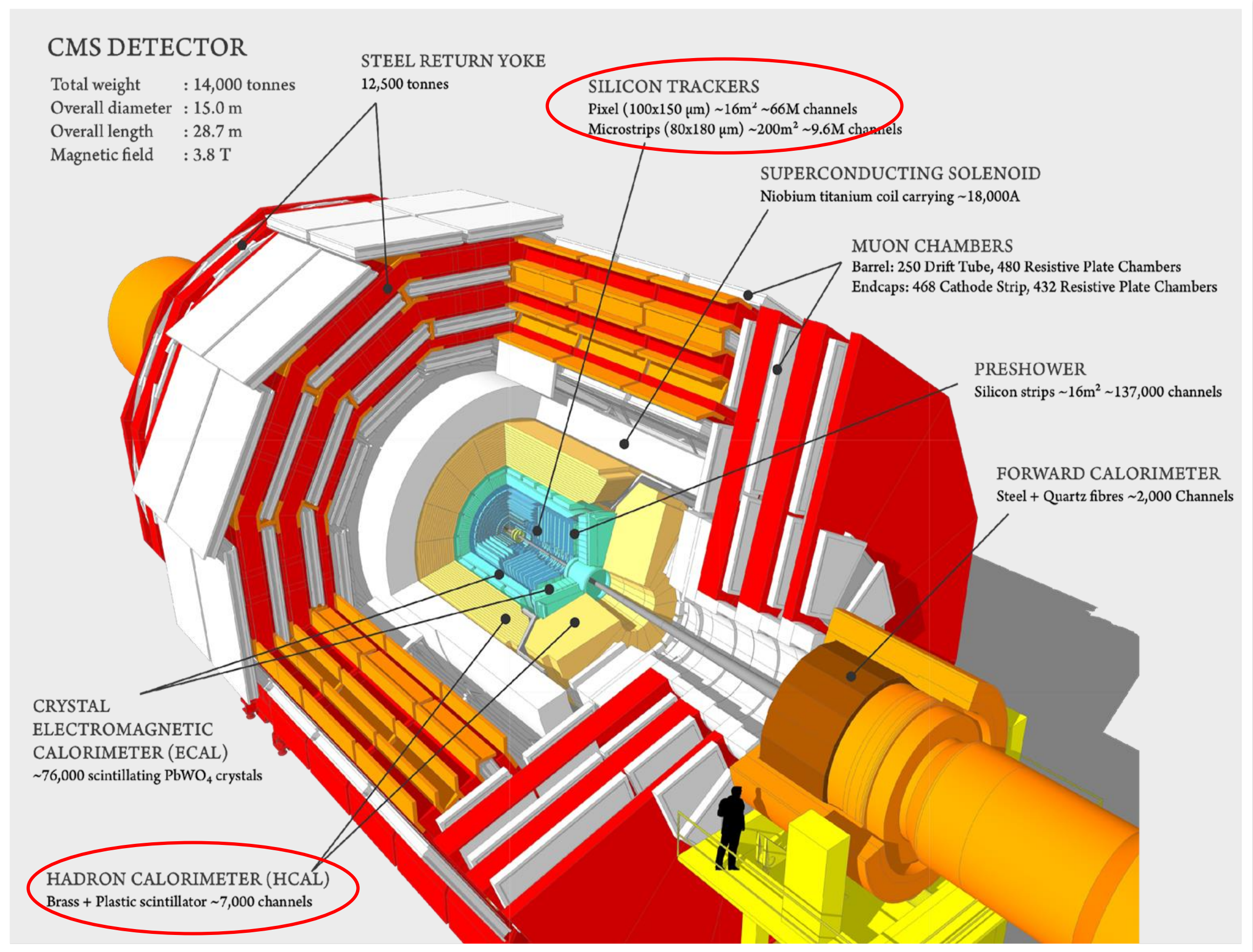
- Provides feedback of live data taking.
- Alarms if something goes wrong.

Offline DQM

- After data taking
- Responsible for bookkeeping and certifying the final data with fine time granularity.

Objectives

- Apply recent progress in Machine Learning techniques regarding automation of DQM scrutiny.
- To compare the performance of different ML algorithms. To compare fully supervised vs semi-supervised approach.
- Impact the current workflow, make it more efficient and guarantee that the data is useful for physics analysis.
- Reduce manpower to discriminate good and bad data, spot problems, save time examining hundreds of histograms. By building intelligence to analyze data, raise alarms, quick feedback.


CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO4 crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Challenges

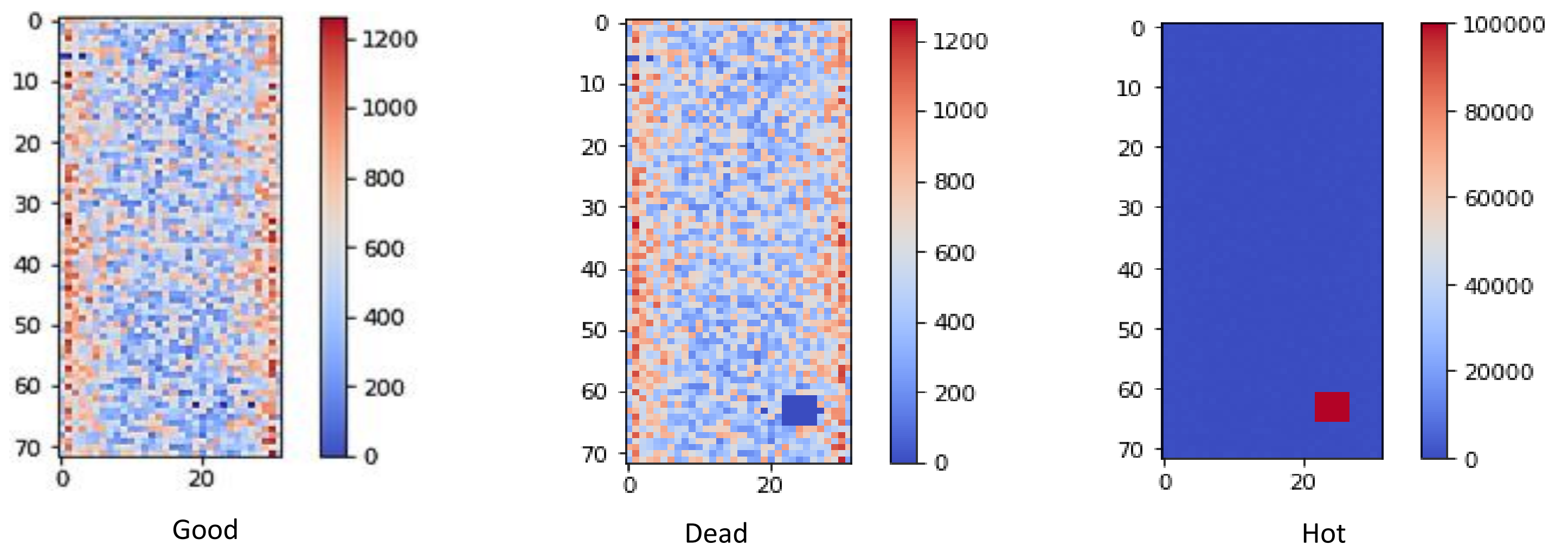
Make sure detector behaves well to perform sensible data analysis.

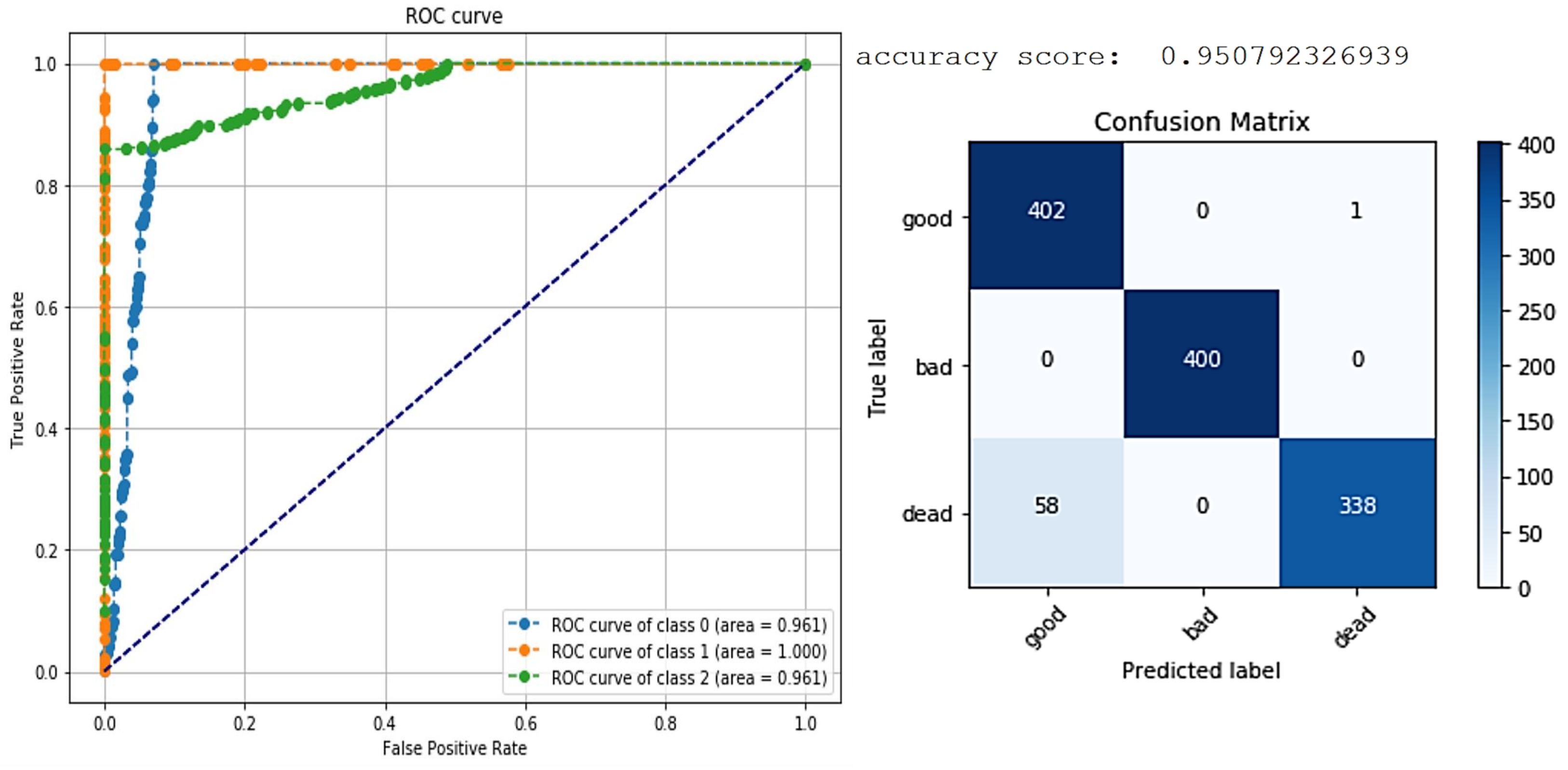
Implementing the best architecture for neural networks

- Underfitting - Too simple and not able to learn
- Overfitting - Too complex and learns very specific and/or unnecessary features

There is no rule of thumb for an ideal model . Many possible combinations.

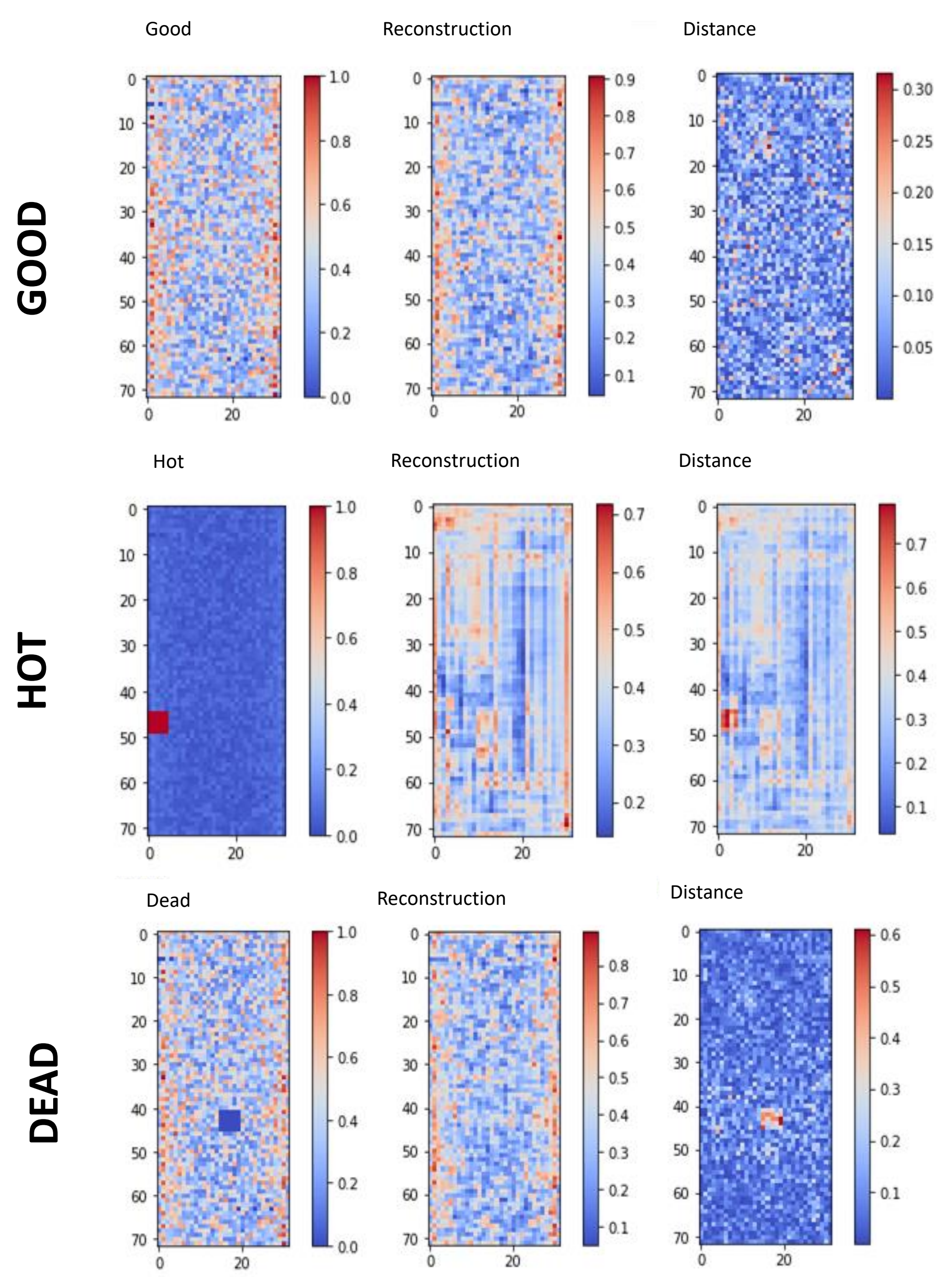
Images used as inputs for ML algorithm (Barrel HCAL) and Performance for Supervised Model

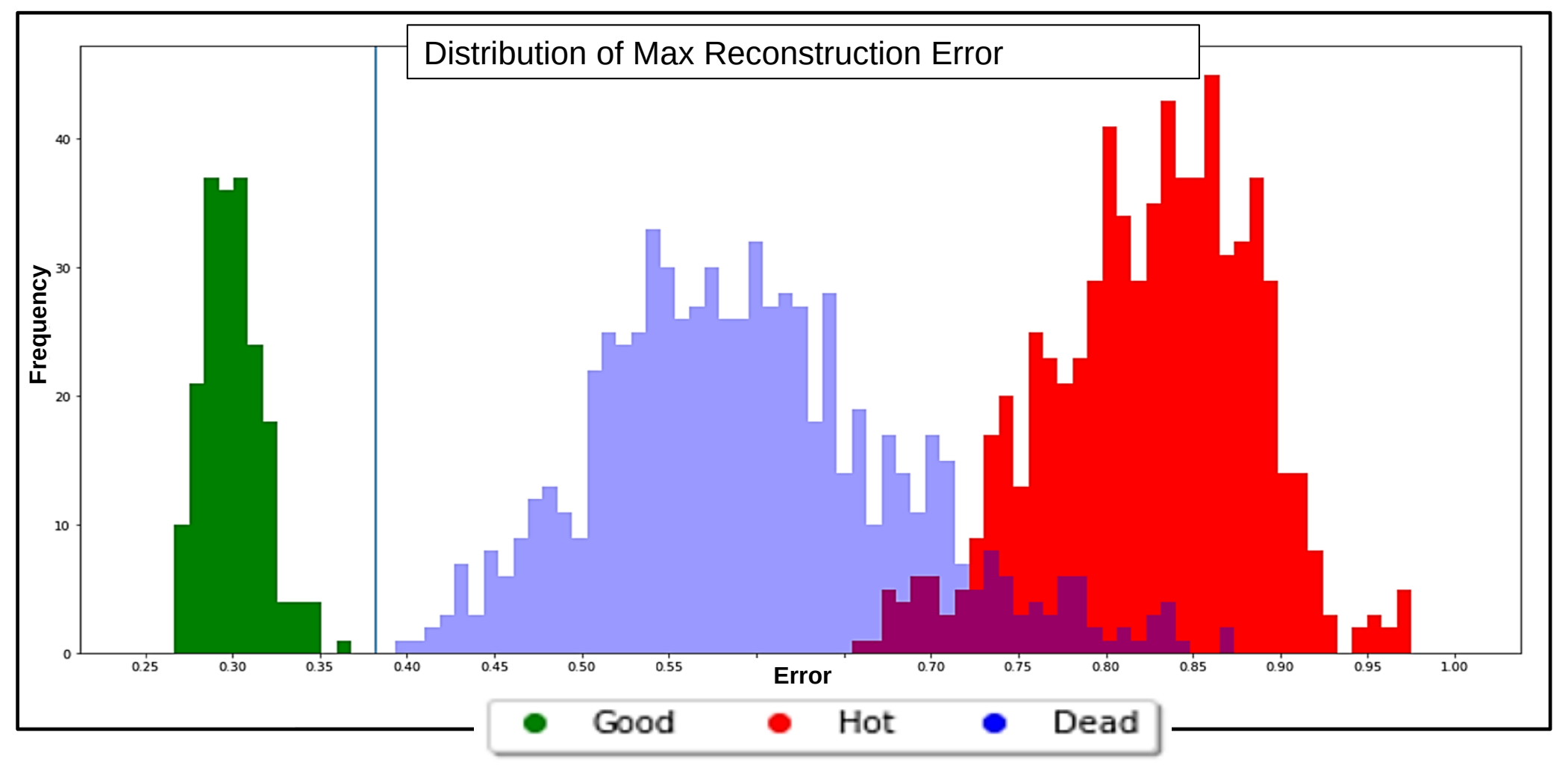


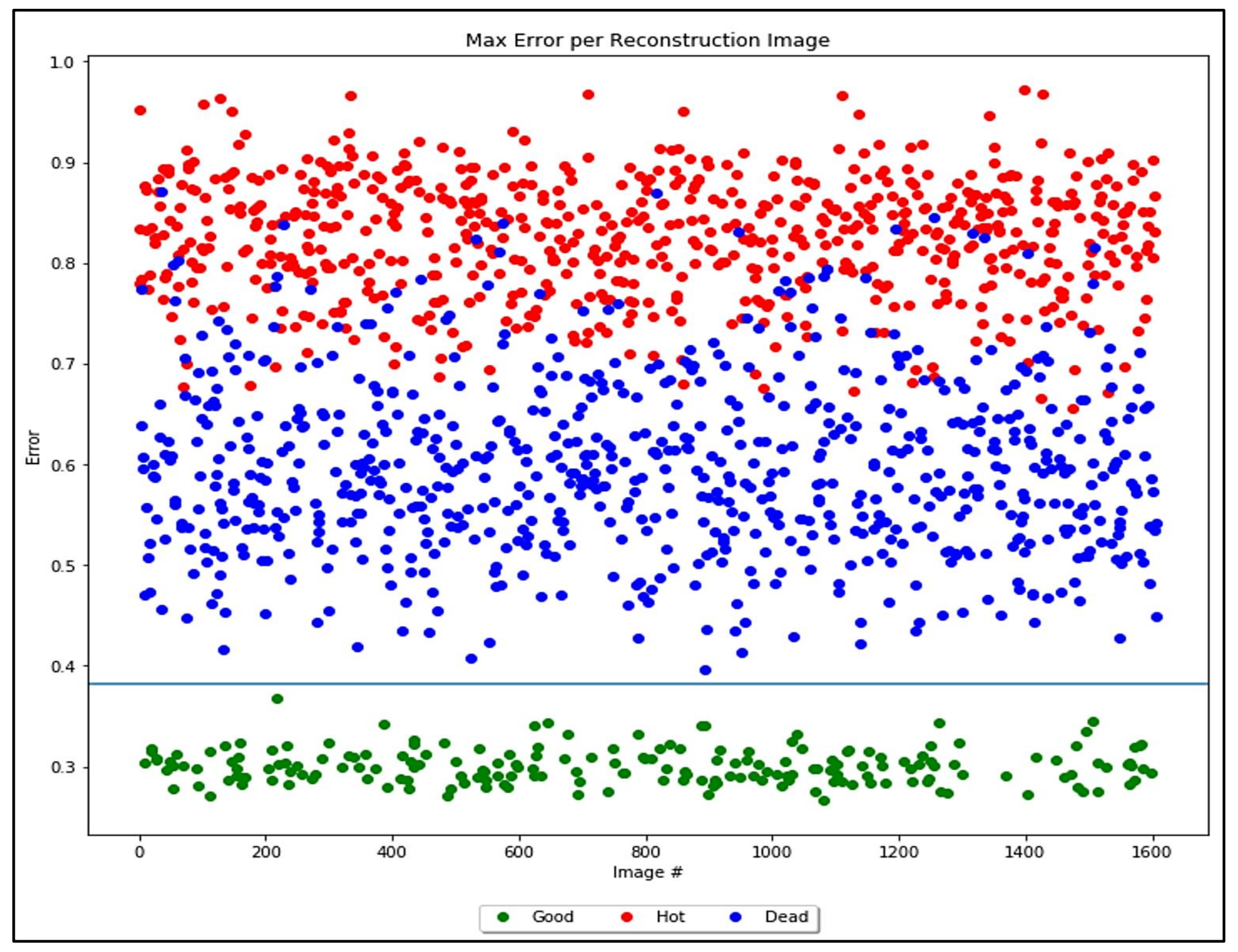

accuracy score: 0.950792326939

- We can see that the model has good performance when discriminating between image classes

Unsupervised Learning Performance







Conclusion

- We can use a Convolutional Neural Network to correctly identify what images fall within the acceptable range to classify as "good".
- We are able to produce an Auto-Encoder that is able to differentiate images that deviate from a baseline image
- Currently, I am using this experience to design DQM using Machine Learning for CMS Silicon Tracker for Phase-2 Upgrade

Acknowledgments

- The US State Dept.
- CERN/CMS
- Prof. Sudhir Malik (UPRM)
- Prof. Steven Goldfarb (UMICH)
- Ph.D Federico de Guio (Texas Tech)
- Society of Physics Students
- Prof. Jean Krisch (UMICH)
- Prof. Nural Akchurin (Texas Tech)